Recent Advances in Robotic Navigation via Large Language Models

Haotian Pan[1], Shibo Huang[1], Jian Yang[2], Jinpeng Mi[3], Ke Li[2], Xiong You[2], Xuan Tang[1], Peidong Liang[4], Jinbo Yang[3], Yingjie Liu[1], Jianfeng Zhang[1], Muyu Wang[1], Jie Yang[1], Xinyu Zhang[1], Lijun Zhao[5], Mingsong Chen[1], Jie Zhou[6], Xian Wei[1] *†

March 6, 2025

Abstract

Recently, with the advances in Large Language Models (LLMs), robot navigation models have demonstrated superior generalization capabilities, including environment perception, decision-making, reasoning, planning, instruction understanding, and human-robot interaction. In this paper, we systematically review recent LLM-based robot navigation research papers, categorizing existing studies into a novel taxonomy comprising perception, planning, control, interaction, and coordination. We also present an overview of the principal datasets and metrics used in robot navigation, analyzing the distinctive characteristics of the datasets and the performance of the main LLMs-based methods. Furthermore, we discuss the challenges hindering the integration of LLMs into robot navigation and provide opportunities and potential directions for future development.

keywords Large Language Models (LLMs), robot navigation, environment perception, decision-making, reasoning, planning, instruction understanding, and human-robot interaction.

1 Introduction

Robot navigation refers to a robot's capability to identify its location within its environment and plan a path to reach a target destination. It is a multidisciplinary field encompassing various aspects such as artificial intelligence, machine learning, computer vision, sensor technology, and robotics. Specifically, the problem of robot navigation is typically regarded as a geometric mapping and planning problem [56]. This implies that robots need to parameterize geometric problems to identify and plan paths from a starting point to a destination in known or unknown environments.

From early model-based approaches to recent advancements in deep learning and reinforcement learning methods, significant progress has been made in robot navigation technology [29, 38, 49, 52, 104]. For example, Leonard et al. [55] utilized the Extended Kalman Filter for the navigation of mobile robots in known environments, and Hu et al. [39] employed landmark identification and subsequent recognition of dynamically extracted environmental features or objects for navigation. With advancements in technology, researchers have started incorporating machine learning to go beyond simple geometric abstractions. These systems are capable of making decisions based on real-world experiences, considering the physical consequences of their actions, and lever-

*[1]Software Engineering Institute, East China Normal University, [2]School of Geospatial Information, Information Engineering University, [3]University of Shanghai for Science and Technology, [4]Fujian (Quanzhou) Institute of Advanced Manufacturing Technology, [5]State Key Laboratory of Robotics and System, Harbin Institute of Technology, China, [6]School of Computer Science and Technology, East China Normal University.

†Corresponding authors: Xian Wei (xwei@sei.ecnu.edu.cn), Jie Zhou, Jian Yang, Jinpeng Mi and Xuan Tang.

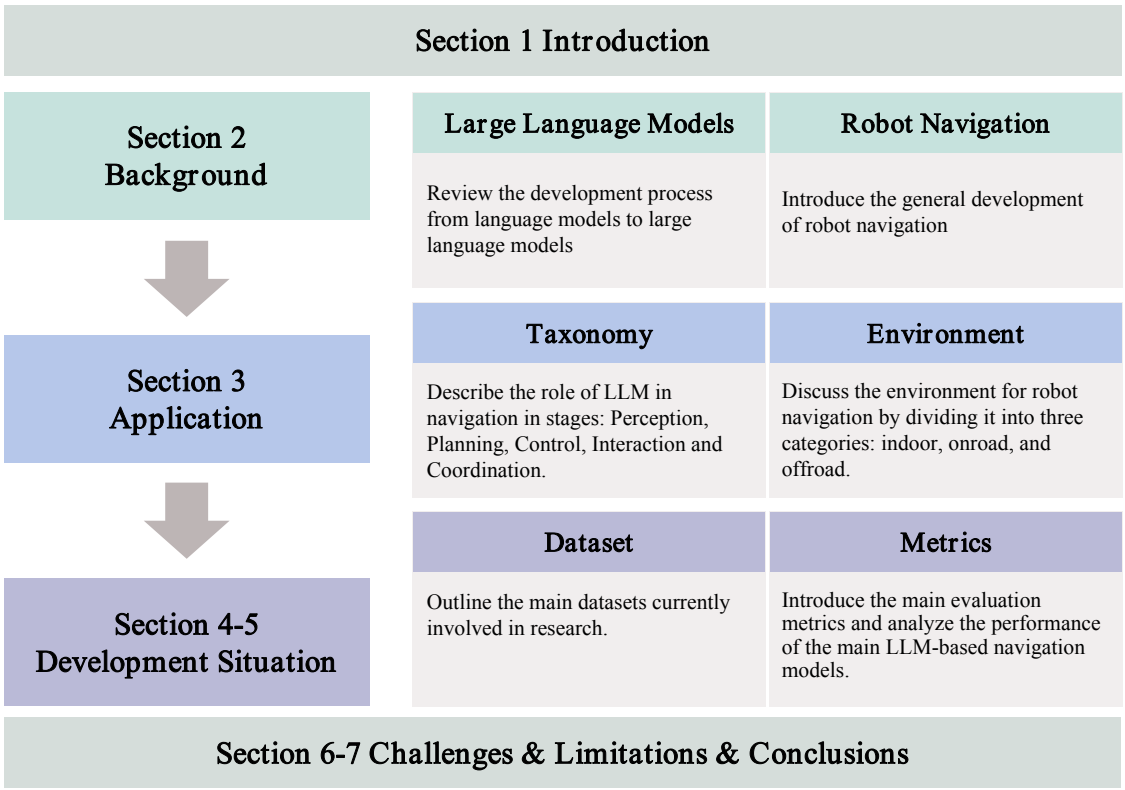

Figure 1: An overview of the main structure of this article.

aging patterns in the real-world environment [96]. However, these methods are data-intensive and lack interpretability, which makes further debugging and improvements challenging. Therefore, many machine learning-based methods are predominantly researched in simulations and only occasionally applied in simple real-world environments as "concept validation" systems for learning navigation [123].

Recently, Large Language Models (LLMs)-based robot navigation methods have garnered significant attention. LLMs such as GPT-3 [14] and BERT [26] are pre-trained on vast amounts of textual data, enabling them to learn rich language patterns, allowing them to perform various language tasks with just a few examples. These models enhance the complexity of embodied intelligent systems with environmental awareness and decision support. Leveraging their powerful language and image processing capabilities, LLMs can effectively plan and make decisions for new tasks with minimal or even without any sample data. LLMs can also be utilized to enhance human-machine interaction. For instance, the LIM2N [146] framework enables language and hand-drawn inputs to serve as navigation constraints and control objectives, showcasing the applications of LLMs in the field of navigation. Furthermore, the concept of generative agents (computational software agents simulating human behavior using LLMs) offers a new perspective for improving human-machine interaction [83].

In summary, the application of LLMs in robot navigation is a promising research direction. However, this field still faces numerous challenges, such as how to effectively encode environmental information into text, how to enable robots to understand and pro-

cess complex environmental information, how to facilitate robots in making rational decisions, how to improve human-robot interaction, and how to achieve autonomous decision-making and reasoning.

To comprehensively understand LLMs-based navigation technology and advance further research in this field, the paper summarizes the latest advancements in LLM-based navigation and discusses future research directions. It is also noted that a recent survey [59, 113] reviewed research related to LLM-based navigation. In comparison to these works, this paper differs in the following aspects:

- This research study focuses on the exploration of LLM-based navigation, which plays a pivotal role in advancing this technology.
- This paper primarily examines the role of LLMs in various aspects of navigation: Perception, Planning, Control, Interaction, and Coordination.
- LLM-based navigation methods are classified based on the physical environments where robot navigation tasks are applied: indoor, on-road, and off-road environments.

This article presents a comprehensive survey of LLMs applied to navigation, encompassing theoretical foundations and practical applications. Figure 1 offers a description of the structure of this work. After the introduction, this paper is divided into six sections. Firstly, in Section 2, we briefly introduce the development of LLMs and robot navigation. Secondly, by examining the different motivations and technologies behind the development of LLMs and robot navigation, we discuss the application of LLMs in the various stages of navigation in Section 3. Next, we study the impact of different models on robot navigation tasks and classify them based on their focus on specific tasks. These efforts are crucial for understanding the connection between LLMs and robot navigation. Thirdly, Section 4 provides a brief comparison and analysis of the relevant datasets. Fourthly, in Section 5, we introduce metrics for evaluating dataset tasks and compare the performance of relevant models. Subsequently, Section 6 discusses unresolved issues, potential research directions, as well as future challenges and research trends. Finally, we conclude the paper in Section 7.

2 Background

In this section, we briefly review the progress in LLMs and robot navigation.

2.1 Large Language Models

LLMs represent a class of Transformer-based language models renowned for their expansive parameter count, often reaching hundreds of billions. These models undergo training using vast quantities of internet data, endowing them with a broad array of language capabilities, primarily manifested through text generation. Prominent examples of LLMs include GPT-3 [14], PaLM [20], LLaMA [107], and GPT-4 [2]. One salient attribute exhibited by LLMs is their emergent proficiencies, such as in-context learning (ICL) [14], instruction comprehension and chain-of-thought reasoning (CoT) [118].

In contrast to traditional machine learning models, the prowess of sizable language models is predominantly evidenced by their deep bidirectional representations, potent context comprehension, and efficient handling of complex tasks. Conventional machine learning models, such as Long Short-Term Memory Networks (LSTM), typically rely on specific data structures and algorithms to process data. On the other hand, substantial language models like GPT-4 and Sora [67], founded on the Transformer architecture, rely solely on attention mechanisms for information processing.

Utilizing Language Model Leveraging to develop embodied models that aid intelligent agents in achieving human-like capabilities represents a burgeoning direction. The language competence and knowledge inherent in LLMs can assist these agents in reasoning about semantic information in the real world and planning executable actions [5, 27, 42, 43, 58, 112]. Through embodied language models [27], data from sensors can be directly converted into interactive code, thereby facilitating a direct transformation from perception to words. The gap between percep-

tual information and text disappears as real-world sensor data seamlessly integrates with language models. Voyager [112] introduces lifelong learning by incorporating three primary constituents: an autonomous curriculum that encourages exploration, a skill repository to store and retrieve intricate behaviors, and an iterative prompting mechanism to generate code for low-level control. Voxposer leverages LLMs to generate robot trajectories for diverse manipulation tasks, guided by open-ended instructions and object cues [43].

2.2 Robot Navigation

Mobile robot navigation technology [30, 105] has garnered widespread attention due to its comprehensiveness and practicality [109]. Over the years, the research in this field has been fruitful, integrating various algorithms ranging from classical control to machine learning [82]. It involves a multi-layered architecture, encompassing the core aspects of perception, planning, and control. Resolving these three core questions involves a series of key technologies, including environmental perception, autonomous localization, and motion planning.

The environmental perception technology of mobile robots involves using sensors carried by the robot to perceive the surrounding environment and processing the acquired environmental data to obtain specific information about the surroundings (including feature and positional information) [93]. In scenarios where the map of the environment is unknown, and the initial position is uncertain, mobile robots must first rely on the sensors they carry to perceive information about the external environment before proceeding with tasks such as localization, map construction, and path planning. Therefore, environmental perception forms the foundation and crucial aspect of autonomous navigation for mobile robots [57].

The localization issue is a fundamental challenge in achieving the autonomous mobility of mobile robots. Depending on the requirements of the task at hand, the localization matter for mobile robots can be divided into pose tracking, global localization, and kidnapping scenarios. Autonomous localization entails the robot utilizing prior environmental map information, the current estimate of the robot's pose, and sensor observations as input data, which, through certain computations, generate a more accurate estimation of the robot's current pose [50].

Path planning involves the robot, in a real-world setting, integrating prior map information and real-time sensory input of the surrounding dynamic environment to search for a trajectory that connects the starting point and the goal point. This trajectory, under specific criteria, is optimal and ensures the robot can navigate past dynamic obstacles in the environment in real-time, ultimately reaching the target smoothly [1, 10, 31]. Achieving autonomous robot navigation necessitates addressing the three core challenges, which are categorized into Map-Based Navigation [77] and Mapless Navigation [34]. Having an environmental map aids in efficient path planning. However, in many scenarios, the environmental map is initially unknown, leading to the development of mapless navigation and Simultaneous Localization and Mapping (SLAM). Moreover, the quality of map-based navigation is directly influenced by the representation and accuracy of the environmental map. Maps can be metric or topological. In metric maps, detected obstacles, landmarks, and the robot's position are represented relative to a specific reference frame. In recent years, algorithms based on traditional search and sampling methods continue to emerge [12, 45, 102]. Furthermore, learning-based planning methods have attracted the attention of many researchers, including approaches that integrate both conventional and learning paradigms [84]. Traditional search methods encompass visual graph search algorithms and grid map-based search algorithms, while sampling-based algorithms include probabilistic roadmaps and methods based on random search trees. Learning-based methods consist of socially aware motion planning based on reinforcement learning. Hybrid methods that combine traditional, and learning approaches involve learning methods for predicting self-vehicle posture in sampling-based incomplete motion planning tasks [90].

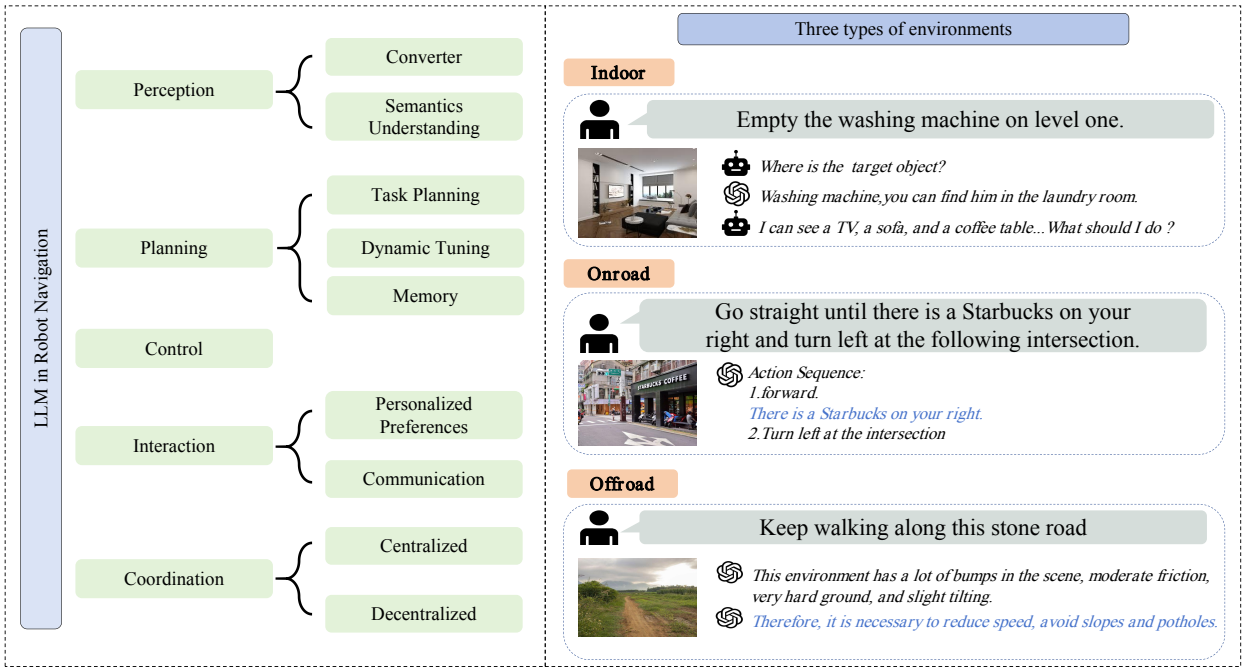

Figure 2: An overview of LLMs for robot navigation.

3 LLMs in Robot Navigation

The development of robot navigation has been swift, yet it remains riddled with a multitude of challenges. The primary technical hurdle lies in path planning. Effective navigation demands a consideration of the dynamic variations within the environment and uncertainties, alongside the art of reaching the intended destination while evading obstacles [4]. As robots navigate through unfamiliar terrains, the uncertainties and intricacies of their surroundings pose additional challenges. Human activities, for instance, exert a noteworthy influence on robot navigation. Robots must adeptly cohabit spaces with humans and avert conflicts [68]. Furthermore, the selection of sensors and the fusion of their data stand out as pivotal challenges in robot navigation. Each sensor boasts distinct levels of precision and reliability—harmonizing the data from diverse sensors to enhance positioning and navigation accuracy stands as a critical avenue of research [61, 113]. To confront these challenges, amalgamating LLMs with robot navigation emerges as a viable convergence.

The fusion of robots with LLMs traces back to 2017, the year that witnessed the inception of the Transformer model founded on attention mechanisms [11]. Subsequently, in 2019, the introduction of the BERT model propelled the realms of deep bidirectional learning representations [25]. These advancements laid a robust foundation for the subsequent fusion of robots with LLMs. Moreover, commencing in 2020, the method of pre-training massive language models and fine-tuning them for specific tasks has led to significant performance enhancements across various NLP tasks [14]. The effective collaboration between robots and LLMs encapsulates a broad spectrum of outcomes enhancing intelligence and human-machine interactions to fortifying autonomy, decision-making prowess, perceptual and control capabilities, and further extending to fostering learning and adaptability, supporting social interac-

tions, and facilitating emotional exchanges, among numerous other dimensions [85, 117, 135, 136]. These breakthroughs not only showcase the extensive application potential of LLMs within the realm of robotics but also furnish novel directions and insights for future research and development endeavors.

In the ensuing sections of this segment, we will delve into the impact of LLMs across various stages of robot navigation and the associated models (Figure 2).

3.1 Robot's Environment Perception and Semantic Understanding

Humans primarily rely on their five senses for perception while moving, without any prior prompts. To endow LLMs with a similar capability, it is necessary to provide them with a comprehensive multimodal description of the environment. To achieve this, a Converter with multimodal perception capabilities is indispensable, as it assumes the responsibility of conveying detailed descriptions of environmental features to the LLMs. Currently, there are three main forms of mainstream Converters:

- **Capturing the correlation between vision and language**, as exemplified in LM-Nav [95], involves utilizing Vision Language Models (VLM) [25] like CLIP as grounding modules for LLMs. The LLMs leverage their semantic understanding capabilities to parse free-form textual instructions and then match the landmark descriptions in these instructions with the images observed by the robot. VLMs, such as CLIP, may not be flawless in this task; for instance, they might fail to detect specific landmarks like fire hydrants or cement mixers. If a landmark exists in the environment and can be recognized by the VLM [6, 24, 43], the robot can effectively localize and follow the landmark. However, if the VLM fails to detect a particular landmark correctly, the robot may choose an incorrect path.

- **Describing multimodal information in natural language**, as exemplified by Matcha [140], involves utilizing Multimodal Perception Modules to convert the results into natural language form for improved comprehension and unified processing by language models. For instance, the visual perception module uses a pre-trained ViLD [36] model to detect object categories and positions in the scene, subsequently generating a scene description. The impact sound perception module categorizes impact sounds and translates them into natural language descriptions. Similarly, weight measurement information is directly transformed into natural language expressions such as "it is light" or "it weighs 30 grams", facilitating seamless integration with language processing frameworks.

- **Encoding original images into Visual Tokens**, akin to the Image-oriented Tokenizer in [142], involves techniques like Vector Quantized Generative Adversarial Network (VQ-GAN). VQ-GAN is an image encoding method that decomposes images into discrete visual units, which are then mapped to a visual dictionary. This process transforms image data into a format that can be processed by language models. Subsequently, a lightweight projection module utilizes a trainable projection matrix to map these visual tokens into the same vector space as text embeddings. Through this approach, LLMs can integrate visual and language information, enabling the generation of multimodal responses.

Upon receiving the multimodal information provided by the Converter, the LLMs engage in semantic reasoning by understanding and parsing environmental information, language instructions, and historical records.

For instance, NavGPT [143] utilizes Visual Foundation Models (VFM) to convert visual information of the environment into natural language descriptions. It then combines navigation system principles and Navigation history to infer the current location of the robot. For example, if the LLMs receive visual information describing "refrigerator" and "sink," and previous history indicates passing through the "kitchen," it may infer that the current location is likely in the kitchen. The LLMs make decisions on the following actions based on this information, such

as selecting actions corresponding to the identified viewpoint ID, but do not directly "recognize" rooms; Instead, they make judgments based on contextual information.

In addition to localization, LLMs can infer interactive information in the environment through perception. For instance, to facilitate effective interaction with objects, as discussed in [106], robots can utilize LLMs to understand the actions each object affords and to determine which objects require avoidance and caution based on language instructions, known as the objects' affordances and constraints [9]. Within the VoxPoser [43] framework, affordances represent how objects can be used, such as a drawer being opened, while constraints limit the possibilities of actions, such as avoiding contact with a vase. By mapping affordances and constraints onto a 3D value map, robots can synthesize motion trajectories to execute complex tasks, taking into account obstacles and objectives in the environment. This approach enables robots to comprehend language instructions and act accordingly, generalize to unseen instructions and attributes with zero-shot learning, without requiring extensive robot-specific data.

3.2 High-level Planning

Robot navigation planning faces challenges across multiple levels, spanning from agile low-level control to high-level planning and reasoning. It requires extensive domain knowledge, and the search space is vast [40, 101]. The goal of navigation planning is to convert natural language instructions, including spatial and temporal constraints, into a set of coordinates or encodings of time path points, such as (x_i, y_i, z_i) [17]. Through vast amounts of pre-training data and self-supervised learning techniques, LLMs can be utilized to provide navigation planning for executing complex long-horizon robot tasks [32]. A prompting approach is introduced in [100], which utilizes LLMs to directly generate action sequences without the need for additional domain knowledge. The empowerment of LLMs for navigation planning tasks includes decomposing instructions into subtasks, tracking navigation progress, and adapting to exceptions in plan adjustments [143] (Figure 3).

Sub-task decomposition involves generating a sequence of subtasks from a given language instruction, where each sub-task is an independent unit that contributes to the overall task. For instance, if a task is "move from point A to point B and then turn back", the subtasks may include "navigate to point B" and "turn back". The decomposed subtasks are handed over to the low-level controller, primarily responsible for executing simple actions and basic environmental interactions [22]. Its objective is to translate each obtained sub-task into a series of executable control commands . For instance, to move from the current position to the table in the current room, the agent's position and orientation are taken as input, and output commands such as forward, left turn, and right turn are generated to control the robot's actions. Existing methods for subtask decomposition can be categorized into three types: Zero-shot Planning, Recursive Planning, and Task Planning + Feedback [17].

- **Zero-shot Planning**, as proposed in [42], involves LLMs generating the entire sub-task sequence at once by integrating geospatial data and natural language instructions, without verifying their executability.
- **Recursive Planning** involves iteratively prompting the LLMs to generate each subsequent subtask in the sequence based on the preceding subtasks. Taking SayCan [5] as an example, the value function module in SayCan generates a value function space based on the current scene and the result of the previous task, representing the likelihood of different task executions. It then selects the most probable next subtask from the candidate subtasks.
- **Task Planning + Feedback** is the third method that combines subtask sequence generation with tracking navigation progress and feasibility checking. It can find a subtask sequence that satisfies the entire task and verify its feasibility before execution. This approach enables the LLMs to have a comprehensive understanding of the navigation history during the navigation process, allowing for basic progress tracking during navigation. For any infeasible subtasks, the LLMs can prompt feedback

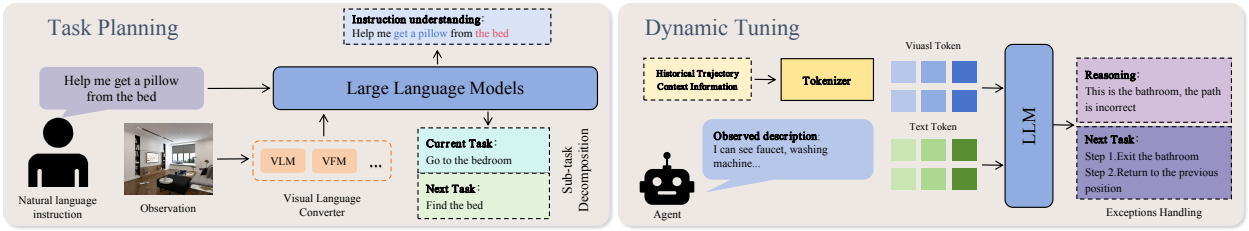

Figure 3: The semantic understanding and contextual reasoning abilities of LLMs help agents decompose instructions into subtasks, track navigation progress, and adapt to abnormal situations in plan adjustments.

regarding the infeasible actions to generate a new subtask sequence. This is similar to the hierarchical approach proposed in [60].

When navigation planning tasks involve complex environments and temporal constraints, simultaneously performing task inference and motion planning can render the decomposition of the problem into subtasks infeasible. To address this challenge effectively, [17] proposed a method that does not directly plan task subtasks using LLMs. Instead, it executes a few-shot transformation from natural language task descriptions to intermediate task representations and then utilizes the TAMP algorithm to jointly solve task and motion planning. Additionally, by introducing self-regressive prompting techniques, it detects and corrects potential synthetic or semantic errors, ensuring that the robot completes tasks as intended.

Furthermore, navigating in unknown environments makes it infeasible to generate a multi-step long-range plan for the assigned task at the outset. This can result in the agent lacking the ability to modify the plan during task execution. During exploration, an effective new location navigation search plan needs to be generated and updated incrementally. For this purpose, the robot needs to use LLMs as a memory to store previously visited areas [69].

As described in SayNav [91], SayNav utilizes a 3D scene graph to support memory for future planning. For instance, it automatically annotates nodes of rooms that have been explored, so if the agent revisits the same room, it does not plan for that room again. Additionally, LLMs are capable of tracking their plans. SayNav achieves its memory by using the conversation chain module in the LangChain framework. A key advantage of this framework is that it avoids reaching the maximum prompt limit on LLMs by not relying on the entire history of the conversation, as is typically done.

Traditional robot memory implementations mainly rely on advanced storage and processing technologies that can simulate human or animal brain memory mechanisms. The memory implementation in robot navigation involves a variety of technologies and methods, ranging from sparsely distributed memory [76] to lifelong learning frameworks [62], as well as experience-based predictive mapping [46] and biologically inspired cognitive models based on hippocampal mechanisms [134]. These methods have their advantages, enabling robots to effectively learn and adapt in different environments, thereby enhancing their navigation capabilities. Tracking based on LLMs may potentially generalize better to other tasks as it eliminates the need for modules that track planning history, thus exhibiting strong generalization capabilities for diverse task implementations.

3.3 Low-level Control

Traditional robot control systems typically employ a series of predefined functions, behaviors, and algorithms to achieve navigation. They use PID controllers to adjust the robot's speed and direction, ensuring that the robot can move to the target location safely and effectively. During this process, hardware limitations and safety constraints are often hard-coded into the control system to prevent unexpected behaviors. To enable LLMs to control

robots, one can facilitate this by teaching LLMs the mapping between natural language and program code. This entails issuing commands or articulating task specifications in natural language, which are subsequently converted into machine-executable program code by the LLMs. For instance, in Code-as-Policies [58], programming-oriented LLMs generate robot policy code based on natural language instructions. Using Large Language Models (LLMs) as controllers for robots offers several advantages. Firstly, LLMs can generate highly flexible robot strategies, aiding robots in adapting to various environments. Secondly, there is no need to collect training data or undergo a training process; instead, existing visual perception models can be directly utilized. As a result, any improvements in the underlying models directly translate to enhanced operational accuracy without additional costs. Lastly, LLM strategy codes exhibit a high level of interpretability. The typical approaches [41, 111] involve designing guiding prompts for LLMs generation. These prompts include Application Programming Interfaces (APIs) and contextual examples. Natural language instructions are then given to the LLMs. In other words, the final input to the LLMs consists of a series of code APIs, usage examples, and task instructions.

Simultaneously, to break the constraint of being limited to using a fixed set of skills for operations, [72] introduces a framework that utilizes an LLM-based planner to query for new skills. This method instantiates skills from an imitation learning agent and, when necessary, gathers human demonstrations to effectively learn new skills. Additionally, in [88], it is also mentioned that combining LLMs with lifelong robot learning allows LLMs to request new skills to accomplish given tasks. For instance, in a scenario where a robot is unable to accomplish a task using existing skill sets, LLMs would request to learn a new skill. By observing demonstration actions provided by a human, the LLMs would acquire this skill and incorporate it into the skill library. This newly acquired skill can be reused in future tasks, demonstrating the potential of open-world learning and lifelong learning [127].

3.4 Human-computer Interaction

Human-robot interaction is crucial in the field of robot navigation, as it enables seamless collaboration and communication between human users and robots. The capability of linking vision with language in VLM can facilitate this interaction. However, the VLM's fuzzy semantic understanding makes human-robot interaction challenging. For example, VLM features may identify a grass field as green and belonging to the grass category but may not infer that it could be the soccer field indicated in a language instruction [24]. Therefore, the application of LLMs in the human-robot interaction module for navigation is essential. LLMs can overcome the limitations of VLM by enhancing the object understanding to achieve more comprehensive and intelligent processing of the environment. Additionally, converting sound into language through voice sensors enables robots to engage in conversational interactions with humans. Through this interaction, robots can update semantic maps [24], eliminate semantic ambiguities, detect ground types [98], and perform functions that cannot be achieved solely through visual information.

The presence of multimodal sensors further assists LLMs in effectively guiding robot behavior in real-world scenarios. The multimodal input forms enable robots to have a more comprehensive and detailed understanding of human commands in real-world settings. For example, as mentioned in LIM2N [146], this framework combines language and hand-drawn sketches to achieve more intuitive and user-centered interactions. Users can guide robots through language instructions or hand-drawn paths, such as drawing paths to guide vehicles. Multimodal inputs can supplement information that is difficult to describe in words. Zhong et al. [79] proposed an intuitive remote operation system using VR devices, demonstrating the feasibility of VR-based control methods. HRC [63] integrates VR thoroughly into robot navigation, utilizing a VR-based remote operation system as a bridge for human-robot collaboration, allowing human supervisors to guide robots through VR devices. This approach enhances the transparency and controllability of remote operations.

In addition to conventional Visual-Linguistic Navigation (VLN), and Socially Aware Navigation (SAN) tasks, navigating in human-populated spaces is a crucial aspect for robots to support various advanced services, such as collaborating with users or walking alongside them. Socially Aware Navigation (SAN) tasks face two main challenges: real-time highly volatile user requests or perceptions, and managing constraints on socially compatible or acceptable navigation behaviors in dynamic environments. With the advancement of robotics and artificial intelligence technologies, current approaches have addressed the challenges of realizing social robots in public environments [133]. These methods draw inspiration from important insights in fields such as machine learning [116], sociology [74], analytical mechanics [73], algebra, geometry [71, 108], and others.

3.5 Multi-robot Coordination

Studying multi-robot collaboration across different domains can help alleviate the limitations of single-agent robot systems, including active mapping [128], exploration [132], and search [66]. In exploration tasks, Multi-Agent Reinforcement Learning (MARL) [131] and Graph-based Learning [128] are both utilized to facilitate the transition of planners from single-agent configurations [66]. Emphasis is placed on Multi-Agent Visual Semantic Navigation, leveraging scene prior knowledge to localize objects in maps and subsequently formulating navigation strategies through reinforcement learning. While both planning-based and learning-based techniques have been successful in multi-robot tasks, they often require common-sense learning in real-world scenarios for robot missions [130]. In contrast, LLMs encompass a broader and richer set of prior knowledge, enabling their application in multi-robot navigation tasks. Recent works [70,137,138] have demonstrated the feasibility of converting environmental observations into linguistic inputs for LLMs to facilitate communication and decision-making within Multi-Agent Systems (MAS). Most works are hierarchical in structure to ensure the smooth operation of MAS. The mainstream LLM-based multi-agent planning frameworks can be broadly categorized into two main branches: centralized [3,18,130,141] and decentralized [64, 114, 129, 138].

- **In a centralized system**, LLMs comprehend the observations, history, and task progress of multiple agents, and collaboratively allocate tasks to each robot group [141] or individual [3]. Taking [130] as an example, the work implements a centralized multi-agent navigation framework that extracts boundary and semantic information from maps and utilizes LLMs to allocate exploration areas to each robot. This framework demonstrates good coordination and planning performance in small-scale teams. However, as the team size grows, the increasing communication and information processing burden of centralized leadership poses challenges to rational and timely planning [18].
- **In a decentralized system**, each robot serves as an autonomous entity that exchanges historical observation results through human-like language communication and makes adaptive decisions [129]. In particular, CoELA [138] provides a systematic template for decentralized communication and collaboration. This approach divides the execution of each agent in a MAS into five modules: observation, belief, communication, reasoning, and planning, where LLMs facilitate communication and reasoning among agents.

3.6 Summary

From existing research, we can observe that the development of robot navigation technology spans from simple geometric feature tracking [55] to complex path planning in dynamic environments. The advancement of these technologies not only enhances the autonomy and efficiency of robots in known or unknown environments but also strengthens their adaptability in complex and dynamic environments. In different environments, robot navigation technologies face various challenges and requirements. The application of LLMs in robot navigation also leads to significant differences in navigation strategies due to the three distinct environments: indoor, on-road, and off-road. These differences are primarily manifested in environmental perception, the selection of

path-planning algorithms, and the application of localization techniques. The discussion will be divided into indoor, on-road, and off-road environments.

- **Indoor environments**. Target objects are more likely to appear near specific rooms and objects, which aids agents in searching for target objects. Therefore, upon detecting room and object information in the current scene, the agent can leverage pre-trained LLMs to perform commonsense reasoning on target objects and semantic scene information through text prompts [144].

- **Onroad environments**. The focus of robot navigation tasks primarily lies in the complexity of the environment and the scale of the navigation task. Involving dense city street networks often requires dealing with more intricate intersections and directional changes, such as 3-way, 4-way, and 5-way intersections, which pose higher demands on landmark detection and directional understanding. Additionally, urban navigation paths typically average 40 steps, far exceeding the average of 6 steps in indoor navigation tasks. Therefore, road environment navigation necessitates continuous visual and language comprehension throughout the entire navigation process [94]. The semantic understanding capability of LLMs can assist the agent in better-identifying landmarks, understanding instructions, and learning trajectory memories. Furthermore, road navigation also needs to consider the influence of human behavior, especially in densely populated urban environments. The ability of LLMs to track user language feedback and adjust robot behavior inference is crucial in road navigation. By capturing human intent and spatiotemporal dependencies, LLMs can comprehend and adapt socially acceptable navigation behaviors in complex dynamic environments [115].

- **Offroad environments**. The variability of terrain significantly impacts robot movement, making it challenging to devise a universally effective navigation policy across all scenarios [47, 124]. LLM-based navigation approaches combine human insights with technical solutions. Humans can intuitively grasp environmental contexts, such as associating wet grass with high impedance, a concept that current algorithms struggle to comprehend but can be elucidated by using LLMs to interpret environmental backgrounds. Furthermore, when robots receive additional contextual information from human observers, such as "you are entering a grassy area after rain" or "you are walking on dry rocky paths under the sun", an LLM-based translator can extract embeddings representing contextual information from human explanations, enhancing the decision-making process in navigation and the robot's sensory observations [98].

In areas such as robot navigation and autonomous driving, numerous models are constantly emerging, each with its unique advantages and applicable scenarios. However, to achieve more efficient and accurate performance, it is of paramount importance to compare and analyze these models. By conducting in-depth research on the capabilities of different models in handling complex environments and completing specific tasks, we can better understand their characteristics, thereby providing more targeted choices and optimization solutions for practical applications.

As shown in Table 1, our comparison of these typical models reveals their strengths and weaknesses, providing important references for future research and development. We should make full use of the advantages of these models and carry out innovation and improvement in combination with actual needs to promote greater breakthroughs in related fields. It is believed that through continuous exploration and practice, these models will play a more significant role in their respective application domains, bringing more convenience and progress to people's lives and the development of society.

4 Datasets

LLMs play a crucial role in assisting robots with navigation. They can accurately understand and parse various instruction information, providing accurate navigation path planning and decision support for robots, greatly enhancing the ability of robots to navigate autonomously in complex environments. At

Table 1: This table summarizes the current mainstream LLM-based navigation methods, which mainly cover the application directions, usage environments, and the role of LLM in the model.

Methods	Environment	The role of LLMs	Application
A2Nav	Indoor	Instruction Parser	Zero-Shot Vision-and-Language Navigation
ESC	Indoor	Zero-shot Object Recognizer	Zero-Shot Vision-and-Language Navigation
NavGPT	Indoor	Path Planner,Controller	Zero-Shot Vision-and-Language Navigation
L3MVN	Indoor	Zero-shot Object Recognizer	Zero-Shot Vision-and-Language Navigation
ATLAS	Indoor	Path Planner	Zero-Shot Vision-and-Language Navigation
MIC	Indoor	Path Planner	Vision and Language Navigation Based on Scene Perception
SayNav	Indoor	Path Planner	Multi-Object Navigation
TaPA	Indoor	Path Planner	Complex Embodied Task Planning
SQA3D	Indoor	Environmental Information Converter	Embodied Scene Understanding
DynaCon	Indoor	Contextual Awareness	Context-Aware Navigation
Co - NavGPT	Indoor	Global Planner for Multiple Robots	Multi robot collaborative navigation
DriVLMe	Onroad	Path Planner	Autonomous Driving Navigation
VELMA	Onroad	Instruction Parser,Path Planner	Vision-and-Language Navigation in street view environment
3P-LLM	Offroad	Path Planner	Path Planning Based on GPT-3.5-Turbo
LM-Nav	Offroad	Instruction Parser	Navigation in complex outdoor environments using free-form text commands
LANCAR	Offroad	Environmental Information Converter	Unstructured terrain navigation

the same time, different datasets also have unique roles in the field of visual navigation of robot language. These datasets cover a rich variety of information, including semantics, visual features, multimodal fusion, and interaction with the environment, providing valuable resources for the learning and training of robots, enabling them to better adapt to various navigation scenarios and achieve more intelligent and efficient navigation.

In Table 2, through a multi-dimensional and meticulous framework, we conduct a profound and all-around sorting and integration of the key datasets in the field of robot language visual navigation. It encompasses aspects such as the scale size, applicable environment, types, and exploration methods of the datasets. It offers a clear, intuitive, and highly valuable reference, facilitating precise comparison and in-depth analysis of the characteristic differences among different datasets in the aforementioned domains.

Semantic information datasets are crucial for improving the understanding and task execution ability of robots. It provides robots with precise semantic knowledge and contextual information, enabling them to accurately understand human instructions and intentions and, thus, better plan and execute tasks. By learning and analyzing this dataset, robots can accurately identify objects, scenes, and concepts, and associate them with corresponding actions and decisions, improving the efficiency of problem-solving and task execution. In addition, it can also enhance the adaptability of robots to different tasks and environments, making them more flexible. The R2R [8] dataset is used to study vision and language navigation, containing $21,567$ natural language instructions, and the agent can navigate in the Matterport3D simulator according to these instructions. The REVERIE [89] dataset contains $21,702$ instructions involving navigation and referential expression information, which can help the agent navigate in a real indoor environment and identify remote target objects. Just Ask [19] proposed to introduce human-computer interaction in the visual and language navigation task based on the R2R dataset to address the ambiguity of the robot, and the proposed framework includes the MC and ASA models and data enhancement strategies to improve the navigation success rate of the robot and make it adapt to the noise in human responses. The FAO [145] dataset

Table 2: This table presents systematically and elaborately the information of numerous datasets in the field of robot language visual navigation. It covers aspects such as the size of the scale, the applicable environment, the type, and the exploration method.

Name	Year	Size			Word		Type			Method	
		Panoramic	Instruction	Object	Domain	Environment	Robot	Vehicles	UAV	Heuristic	Stationary
Matterport3D	2017	10800	-	-	Indoor	Simulator	✓	✗	✗	✗	✓
AI2 - THOR	2017	-	-	3575	Indoor	Simulator	✓	✓	✗	✓	✗
R2R	2018	-	21567	-	Indoor	Matterport3D	✓	✗	✗	✗	✓
Gibson	2018	1447	-	-	Indoor	Real	✓	✓	✓	✓	✗
IQA	2018	-	75000	-	Indoor	AI2 - THOR	✓	✗	✗	✓	✗
EmbodiedQA	2018	-	5281	50	Indoor	House3D	✓	✗	✗	✓	✗
RoomNav	2018	404508	-	5697217	Indoor	House3D	✓	✗	✗	✓	✗
TOUCHDOWN	2019	29641	9326	-	Onroad	Google Street View	✓	✗	✗	✓	✓
CVDN	2019	-	2050	-	Indoor	Matterport3D	✓	✗	✗	✓	✗
VNLA	2019	-	160777	-	Indoor	Matterport3D	✓	✗	✗	✓	✗
HANNA	2019	-	8586	-	Indoor	Matterport3D	✓	✗	✗	✓	✗
StreetNav	2019	580415	-	-	Onroad	Google Street View	✓	✗	✗	✓	✗
StreetLearn	2019	58000	-	-	Onroad	Google Street View	✓	✗	✗	✓	✗
XL - R2R	2019	-	17394	-	Indoor	Matterport3D	✓	✗	✗	✓	✗
Just Ask	2019	-	21567	-	Indoor	Matterport3D	✓	✗	✗	✓	✗
REVERIE	2020	-	21702	-	Indoor	Matterport3D	✓	✗	✗	✗	✓
RobotSlang	2020	-	169	-	Indoor	Real	✓	✗	✗	✓	✗
RxR	2020	-	126000	-	Indoor	Matterport3D	✓	✗	✗	✓	✗
VLN - CE	2020	10800	21567	-	Indoor	Matterport3D	✓	✗	✗	✓	✗
ALFRED	2020	-	25743	84	Indoor	AI2 - THOR	✓	✗	✗	✓	✗
Retouchdown	2020	29641	36901	-	Onroad	Google Street View	✓	✗	✗	✓	✗
HM3D	2021	-	-	1000	Indoor	Real	✓	✗	✗	✓	✗
Robo - VLN	2021	-	9533	-	Indoor	Matterport3D	✓	✗	✗	✓	✗
FAO	2021	-	3848	-	Indoor	Matterport3D	✓	✗	✗	✓	✗
Talk2Nav	2021	43630	10714	20000	Onroad	Google Street View	✓	✗	✗	✓	✗
DialFRED	2022	-	53000	-	Indoor	AI2 - THOR	✓	✗	✗	✓	✗
ProcTHOR	2022	10000	-	-	Indoor	Simulator	✓	✗	✗	✓	✗
TEACh	2022	-	3047	-	Indoor	AI2 - THOR	✓	✗	✗	✓	✗
AerialVLN	2023	-	25338	870	Onroad	Simulator	✗	✗	✓	✓	✗
HM3DSEM	2023	-	-	142646	Indoor	Real	✓	✗	✗	✓	✗
Open X - E	2024	-	160266	23486	Indoor	Real	✓	✗	✗	✓	✗

is based on the Matterport3D environment and provides instructions containing object attributes, relationships, and other information so that the agent can find the target object from any position. The VNLA [81] dataset contains rich indoor environment information and target object instructions, and the agent learns to find the target object through interaction with the advisor and independent exploration. The HANNA [80] dataset is also based on the Matterport3D environment, simulating a scene where an agent looks for objects in an indoor environment, and the agent can request help, and researchers can train the agent's ability to learn to request and interpret language and visual instructions through the indoor environment and interaction records. The XL-R2R [126] dataset is the first cross-language VLN dataset, which extends the R2R dataset and adds Chinese instructions to support the research of cross-language VLN tasks. The ALFRED [99] dataset contains 25,743 natural language instructions for acquiring the ability to translate natural language instructions and egocentric vision into sequences of actions for domestic chores.DialFRED [33] is a conversational embodied instruction-following benchmark extended based on the ALFRED benchmark, allowing the robot to actively ask questions and use the

answer information to complete the tasks when performing household tasks, and the dataset contains related tasks, questions and answers, and oracles. The IQA [35] dataset is based on the AI2-THOR environment and contains more than 75,000 multiple-choice questions, and the agent needs to navigate in the scene, interact with objects, and plan actions to answer the questions. The TEACh [87] dataset contains 3,047 game sessions for completing household tasks in the AI2-THOR simulator, which can be utilized to train and assess the models' ability to complete household tasks. Based on the House3D virtual environment, the RoomNav [120] dataset allows the agent to navigate from a random position to the target room according to the instructions of high-level semantic concepts. The questions in the EQA [21] dataset are represented by executable functional programs to test the various abilities of the agent.

Visual feature datasets are extremely critical for improving the visual perception ability of robots. It provides rich feature information such as object shape, color, texture, etc., enabling robots to more accurately perceive and recognize the surrounding environment. By learning and analyzing this dataset, robots can optimize the visual perception model and improve the processing and understanding ability of visual information. This leads to better navigation, avoiding obstacles, and other operations more accurately, and can also better adapt to different lighting and environmental changes, enhancing the reliability and stability in complex environments. Matterport3D [15] is a large-scale RGB-D dataset containing 10,800 panoramic views of 90 building-scale scenes, composed of 194,400 RGB-D images, and also includes annotations for surface reconstruction, camera poses, and 2D and 3D semantic segmentation, providing support for a variety of computer vision tasks. The TOUCHDOWN [16] dataset builds an outdoor visual street environment based on Google Street View, containing 29,641 panoramas and 61,319 edges, covering New York City, and is used to study natural language navigation and spatial reasoning. The RxR [54] dataset is a new visual and language navigation dataset, and the path design of the dataset meets a variety of expected characteristics, including high variance in path length, indirect approach to the target, naturalness, and uniform coverage of environmental viewpoints, to ensure that language plays a key role in navigation. Multimodal fusion datasets can effectively enhance the ability of robots to process complex information. It integrates multiple modal data, such as text, image, audio, etc., to provide robots with comprehensive and rich information input. By fusing and analyzing these data, robots can better understand the correlation and complementarity of different modalities, more accurately process complex situational information, more flexibly respond to various challenges, and help continuously improve their intelligence level and processing ability. AI2-THOR [51] is an interactive 3D environment framework for visual AI research, containing a variety of scene datasets, agents, supported actions, image modalities, environmental metadata, etc., and interacts with the Unity 3D game engine through the Python API, and the agent can perform a variety of operations in the scene.

Datasets involving interaction with the environment are the fundamental basis for training robots to effectively navigate and execute tasks. Such datasets contain rich environmental information, providing robots with training scenarios that simulate the real environment. By learning these datasets, robots can master the skills of interacting with the environment, improve the accuracy and efficiency of navigation, and can also better understand the task requirements, and make appropriate decisions to ensure the smooth execution of the task, and its role is crucial. The RobotSlang [13] dataset is collected through cooperative experiments in which human drivers control physical robots and human commanders provide navigation guidance and are used to study language-guided robot simultaneous localization and mapping. The CVDN [103] dataset contains 2,050 human navigation dialogues that occur in a simulated home environment and is used to study robot navigation through dialogue in a human environment. The Gibson [122] dataset is a virtual environment for training and testing real-world perception agents, built based on the scanning of real spaces, and has the characteristics of neural network view synthesis and physical engine integration. The PROCTHOR [23]dataset contains 10,000 fully inter-

active houses, including different room layouts, furniture and item placements, lighting and material settings, etc., to train and assess the ability of agents in navigation, interaction, and operation tasks. The HM3D [92] dataset is a large-scale indoor 3D environment dataset composed of 3D reconstructions of 1,000 building-scale from different locations in the real world. The HM3DSEM [125] dataset is the largest 3D real-world spatial dataset with densely annotated semantics currently available in the academic community, adding a dense semantic annotation layer based on the HM3D dataset. The Open X-Embodiment [86] dataset is a large-scale robot learning dataset collected by 21 institutions, containing more than 1 million real robot trajectories from 22 types of robots. The Robo-VLN [44] dataset forms a continuous control form by corresponding the human-annotated instructions in the R2R dataset to the sparse waypoints in the 3D reconstructed environment, which is used to solve robots navigating according to natural language instructions in a continuous environment. The VLN-CE [53] dataset converts the trajectory based on panoramic images in R2R into a fine path in the continuous Matterport3D environment, which is used to study the navigation ability of robots in an environment closer to the real world. The AerialVLN [65] dataset is a task designed for drones to navigate in an urban-level outdoor environment, and the dataset generates flight paths by experienced drone operators and annotates instructions by AMT workers. In the interactive environment built based on Google Street View, the Talk2Nav [110] dataset realizes automatic pathfinding based on language in the outdoor environment by obtaining road nodes and street view images in New York City. The StreetNav [37] dataset provides the agent with a series of tasks similar to humans, following navigation instructions to reach the destination in the city by randomly sampling the starting and target positions. The Retouchdown [75] dataset provides human-annotated instructions and spatial description parsing tasks for navigating the streets of New York City. StreetLearn [78] defines navigation tasks such as courier tasks, and the agent needs to learn the navigation strategy based on visual observation and the encoding of the target location.

This research focuses on the datasets in the field of robot language visual navigation, clarifies the importance of LLMs for robot navigation, and the unique roles of different datasets in semantic information, visual features, multimodal fusion, and environment interaction. Through the classification and analysis of multiple datasets, a clear framework is provided for related research, which helps to promote the improvement of robot navigation ability and is expected to bring innovative breakthroughs in multiple fields in the future and have a positive impact on scientific and technological progress.

5 Evaluation Metrics and Analysis

The navigation techniques based on LLMs have been widely applied in many fields of today's society, making the optimization and improvement of its performance a vitally important research direction.

To abstract the effectiveness of LLM-based navigation into mathematical terms, this paper presents mainstream evaluation metrics. These metrics can provide insights into the accuracy and adaptability of the models, encompassing aspects such as navigation accuracy, efficiency, and the model's adherence to instructions [121].

Among these metrics, the most straightforward one that can represent the target navigation is the Success Rate (SR) [7], which signifies the frequency at which tasks are completed within the specifically defined proximity to the target. The path length (PL) [119] signifies the overall length of the navigation route, and the shortest path distance (SPD) evaluates the final position reached from the initial position to the pre-determined destination. The Trajectory Length (TL) [28] represents the average distance travelled by the robot. Longer paths may reduce the efficiency of navigation, as they lead to increased wear and higher resource utilization for actual robots. To address this issue, success weighted by Path Length (SPL) [7] is utilized to strike a balance between SR and path efficiency. The Oracle Success Rate (OSR) assesses whether any portion of the path is close to the pre-

defined target location. Additionally, there is the Remote Grounding Success Rate weighted Path Length (RGSPL), which reconciles the Remote Grounding Success (RGS) failure rate with the efficiency of the navigation paths. Key Point Accuracy (KPA) gauges the correct decision rate of key points. Finally, Distance to Goal (DTG) can be used to represent the minimum distance between the robot and the target when the episode concludes.

Success Rate (SR) delineates the proportion of tasks completed accurately, signifying the efficiency of achieving the goal. It is computed as follows:

$$SR = \frac{\text{Number of tasks accurately completed}}{\text{Total tasks}} \quad (1)$$

Path Length (PL) mirrors the total distance covered throughout the task completion process. A shorter path signifies greater efficiency. It is defined as follows:

$$PL = \frac{1}{N}\sum_{i=1}^{N} P_i \quad (2)$$

Success weighted by Path Length (SPL) combines SR and PL to evaluate the efficiency of task completion. It is calculated as follows:

$$SPL = \frac{1}{N}\sum_{i=1}^{N} S_i \times \frac{\min(P_i, L_i)}{L_i} \quad (3)$$

where S_i indicates the success of the task i (1 if successful, 0 otherwise), P_i denotes the navigated path length, L_i represents the shortest path, and N is the count of tasks.

Oracle Success Rate (OSR) indicates whether the distance from any point on the navigation path to the target is within a pre-defined threshold. It is determined as follows:

$$OSR = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(\min_{n \in Pi} \text{dist}(n, g_i) \leq \text{Threshold}) \quad (4)$$

The meaning of $\mathbb{I}(\min_{n \in Pi} \text{dist}(n, g_i) \leq \text{Threshold})$ is to return 1 if the minimum distance from any node within the path to the target is less than or equal to the threshold, and otherwise it returns 0.

Under the Remote Grounding Success Rate(RGS) standard, success is only attained when the agent locates an instance that corresponds to the target semantic label. This metric is widely used in goal-oriented tasks. The Remote Grounding Success Rate Weighted by Navigation Path Length (RGSPL) is readily comprehensible as it correlates the efficiency of a robot in achieving Remote Grounding Success with the path length it traverses. It is calculated as follows:

$$RGSPL = \frac{1}{N}\sum_{i=1}^{N} \frac{{S_i}^{RGS}}{\max(P_i, L_i)} \quad (5)$$

Similar to OSR, S_i^{RGS} is an indicator where S_i^{RGS} is 1 when task i is successful, otherwise 0, P_i represents the navigated path length, L_i stands for the shortest path, and N denotes the number of tasks. Subsequently, the paper attempts to analyze LLM-based navigation methods using the same metrics. Unfortunately, the evaluation criteria for current methods are not standardized, with each method having its uniqueness. Therefore, Table 3 integrates performance data of various LLM-related navigation models in different environments using SR and SPL. By comparing and analyzing data from different models in environments such as HM3D, R2R, and MP3D, we can clearly and definitively understand the strengths and weaknesses of each model.

To sum up, the in-depth analysis of the evaluation indicators and performance of navigation models enables us to distinctly perceive the achievements and existing deficiencies of the current research. These research findings provide a solid foundation for the optimization and improvement of existing models and also indicate a clear direction for subsequent research.

6 Challenges and Limitations

Applying LLMs to the field of robot navigation has shown significant potential. However, to ensure comprehensive and rigorous research, the following points are provided to elucidate the obstacles to applying LLMs to Robot Navigation for consideration.

Table 3: This table provides a summary of the performance of various LLM-related navigation models in different environments, encompassing indicators like success rate (SR) and success-weighted path length (SPL). These models are deployed in diverse environments such as HM3D, R2R, and MP3D, and leverage the language understanding capability of LLMs to augment the navigation effect.

Methods	Dataset	SR↑	SPL	DTG	TL	OSR	Construction	LLM
A^2Nav	RxR	16.80	6.30	-	-	-	-	GPT - 3
$A^2$2Nav	R2R	22.60	11.10	-	-	-	-	GPT - 3
ESC	MP3D	28.70	14.20	-	-	-	Visual Observations	Deberta v3
NavGPT	R2R	34.00	29.00	-	11.45	42.00	Visual Observations	GPT - 4
ESC	RoboTHOR	38.10	22.20	-	-	-	Visual Observations	Deberta v3
ESC	HM3D	39.00	22.40	-	-	-	Visual Observations	ChatGPT
ESC	HM3D	39.20	22.30	-	-	-	Visual Observations	Deberta v3
VELMA	Street View	49.10	-	-	-	-	Visual Observations	GPT - 3
L3MVN (Zero-Shot)	HM3D	50.40	23.10	4.427	-	-	Semantic Map	RoBERTa - large
Single - NavGPT	HM3D	53.90	21.50	2.633	-	-	Visual Observations	GPT - 3.5 - turbo
L3MVN (Feed-forward)	HM3D	54.20	25.50	3.934	-	-	Semantic Map	RoBERTa - large
VELMA	Street View	56.10	-	-	-	-	Visual Observations	GPT - 4
MIC	REVERIE	56.97	43.60	-	20.64	62.37	Visual Observations	GPT - 2
SayNav	AI2 - THOR	60.32	34.00	-	-	-	Visual Observations	GPT - 3.5
TaPA	AI2 - THOR	61.11	-	-	-	-	Visual Observations	GPT - 3.5
DynaCon	virtual	62.50	-	-	-	-	Visual Observations	GPT - 3.5
SayNav	AI2 - THOR	64.34	33.00	-	-	-	Visual Observations	GPT - 4
Co - NavGPT	HM3D	66.10	33.10	1.831	-	-	Visual Observations	GPT - 3.5 - turbo
DriVLMe	CARLA	70.80	-	-	-	-	Visual Observations	Vicuna - 7B(v1.1)
SayCan	Real	74.00	-	-	-	-	-	PaLM
Co - NavGPT	HM3D	75.70	44.80	1.131	-	-	Ground Truth	GPT - 3.5 - turbo
L3MVN (Zero-Shot)	Gibson	76.10	37.70	1.101	-	-	Semantic Map	RoBERTa - large
L3MVN (Feed-forward)	Gibson	76.90	38.80	1.008	-	-	Semantic Map	RoBERTa - large
SayNav	AI2 - THOR	80.62	32.00	-	-	-	Ground Truth	GPT - 3.5
3P-LLM	Gazebo	81.00	-	-	-	-	-	GPT - 3.5 - turbo
SayNav	AI2 - THOR	84.09	36.00	-	-	-	Ground Truth	GPT - 4

- **Limitations in spatial reasoning abilities**. Despite excelling in sequence modeling and pattern recognition, LLMs still exhibit shortcomings when it comes to handling complex spatial reasoning tasks. For instance, models like ChatGPT-3.5 struggle with processing 3D robot trajectory data and may require specific prompting mechanisms to enhance performance [97].

- **Adaptation issues in the physical world**. When applying LLMs to robot navigation, a key challenge lies in effectively integrating these models with perception and action control in the physical world. While some research has shown that LLMs can generate low-level control commands with minimal physical environment cues, this remains a complex problem, especially in scenarios requiring dynamic adjustments of robot behavior [117].

- **Compatibility issues in visual perception**. For tasks reliant on intricate environmental interactions, such as robotic visual navigation, relying solely on textual instructions may not suffice to encompass all necessary information. While multimodal LLMs (e.g., GPT-4V) have started being utilized to enhance tasks like robot motion planning [113], this still constrains the application of LLMs in tasks requiring high levels of visual perception.

- **Safety considerations, as embodied platforms acting as agents**. Robots' actions can have lasting impacts on the environment; therefore, they need to learn and interact safely to prevent potentially catastrophic events. This necessitates

that physical navigation systems possess adequate self-monitoring and risk assessment capabilities.

- **Demands for real-time and reliability**. In practical applications, robot navigation systems need to respond swiftly and make accurate decisions in uncertain environments. This necessitates that LLMs not only possess efficient processing capabilities but also reduce time delays while ensuring the correctness and reliability of plan execution [48].

- **Diversity in visual navigation**. Visual navigation encompasses a variety of tasks such as visual-language navigation, grounded question answering, scene navigation, etc. [139] This diversity necessitates that embodied navigation systems exhibit sufficient flexibility and adaptability.

These limitations and challenges present an alternative perspective on the current research progress in LLM-based navigation. The primary challenge at present lies in the grounding mechanism of LLMs. Given a sufficient volume of training data, LLMs can effectively address navigation tasks in large-scale environments. However, this is contingent upon perfect visual perception and excellent robotic control strategies. In current research, LLMs often exist as processors or components within navigation models, particularly in outdoor navigation tasks, limiting their efficiency and success rates in task completion.

On the other hand, to fully leverage the capabilities of LLMs, a vast amount of data is required. Current datasets typically focus on indoor environments or specific settings, making it challenging for researchers to acquire extensive datasets encompassing complex environments. Accelerating progress in this field necessitates researchers to delve deeply into the aforementioned issues.

7 Conclusions

This paper delves into the development of LLMs in the field of robot navigation, thoroughly examining various methods within LLM-based Navigation and describing their differences and commonalities. Furthermore, it provides an in-depth analysis of the performance of these methods across various tasks and environments, revealing their fundamental design principles and approaches. The experimental results underscore the crucial role of LLMs in zero-shot navigation tasks. The paper also outlines the inherent challenges and limitations of LLMs in the domain of robot navigation. These factors include the lack of direct connection between LLMs and the physical world, the need for large amounts of training data, and the complexity of understanding and generating natural language in different environments. Despite these obstacles, there are promising research trajectories that can drive progress in LLM-based Navigation. These measures include developing more robust and adaptive language models, investigating innovative training methods and architectures, establishing standardized benchmarks and evaluation metrics, and the urgent need for interdisciplinary collaboration among researchers in artificial intelligence, robotics, and social sciences.

References

[1] S Haider Abdulredah and D Jasim Kadhim. Developing a real time navigation for the mobile robots at unknown environments. *Indones. J. Electr. Eng. Comput. Sci.(IJEECS)*, 20:500–509, 2020.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[3] Saaket Agashe, Yue Fan, and Xin Eric Wang. Evaluating multi-agent coordination abilities in large language models. *arXiv preprint arXiv:2310.03903*, 2023.

[4] Swati Aggarwal, Kushagra Sharma, and Manisha Priyadarshini. Robot navigation: Review of techniques and research challenges. In *2016 3rd International Conference on Computing for*

Sustainable Global Development (INDIACom), pages 3660–3665. IEEE, 2016.

[5] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

[6] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[7] Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, et al. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*, 2018.

[8] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3674–3683, 2018.

[9] Paola Ardón, Èric Pairet, Katrin S Lohan, Subramanian Ramamoorthy, and Ronald Petrick. Affordances in robotic tasks–a survey. *arXiv preprint arXiv:2004.07400*, 2020.

[10] Ioannis Arvanitakis, Anthony Tzes, and Konstantinos Giannousakis. Synergistic exploration and navigation of mobile robots under pose uncertainty in unknown environments. *International Journal of Advanced Robotic Systems*, 15(1):1729881417750785, 2018.

[11] Vaswani Ashish. Attention is all you need. *Advances in neural information processing systems*, 30:I, 2017.

[12] Ben Beklisi Kwame Ayawli, Ryad Chellali, Albert Yaw Appiah, and Frimpong Kyeremeh. An overview of nature-inspired, conventional, and hybrid methods of autonomous vehicle path planning. *Journal of Advanced Transportation*, 2018(1):8269698, 2018.

[13] Shurjo Banerjee, Jesse Thomason, and Jason Corso. The robotslang benchmark: Dialog-guided robot localization and navigation. In *Conference on Robot Learning*, pages 1384–1393. PMLR, 2021.

[14] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[15] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017.

[16] Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12538–12547, 2019.

[17] Yongchao Chen, Jacob Arkin, Charles Dawson, Yang Zhang, Nicholas Roy, and Chuchu Fan. Autotamp: Autoregressive task and motion planning with llms as translators and checkers. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6695–6702. IEEE, 2024.

[18] Yongchao Chen, Jacob Arkin, Yang Zhang, Nicholas Roy, and Chuchu Fan. Scalable multi-robot collaboration with large language models: Centralized or decentralized systems? In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4311–4317, 2024.

[19] Ta-Chung Chi, Minmin Shen, Mihail Eric, Seokhwan Kim, and Dilek Hakkani-Tur. Just ask: An interactive learning framework for vision and language navigation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 2459–2466, 2020.

[20] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

[21] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–10, 2018.

[22] Thomas Dean. High-level planning and low-level control. In *Intelligent Robots and Computer Vision VI*, volume 848, pages 496–501. SPIE, 1988.

[23] Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Jordi Salvador, Kiana Ehsani, Winson Han, Eric Kolve, Ali Farhadi, Aniruddha Kembhavi, and Roozbeh Mottaghi. Procthor: Large-scale embodied ai using procedural generation. *ArXiv*, abs/2206.06994, 2022.

[24] Yinan Deng, Jiahui Wang, Jingyu Zhao, Xinyu Tian, Guangyan Chen, Yi Yang, and Yufeng Yue. Opengraph: Open-vocabulary hierarchical 3d graph representation in large-scale outdoor environments. *arXiv preprint arXiv:2403.09412*, 2024.

[25] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[26] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding.

[27] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

[28] Stefan Edelkamp and Stefan Schrödl. *Heuristic search: theory and applications*. Elsevier, 2011.

[29] Alberto Elfes. Using occupancy grids for mobile robot perception and navigation. *Computer*, 22(6):46–57, 1989.

[30] Felix Endres, Jürgen Hess, Nikolas Engelhard, Jürgen Sturm, Daniel Cremers, and Wolfram Burgard. An evaluation of the rgb-d slam system. In *2012 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1691–1696. IEEE, 2012.

[31] Torvald Ersson and Xiaoming Hu. Path planning and navigation of mobile robots in unknown environments. In *Proceedings 2001 IEEE/RSJ International Conference on Intelligent Robots and Systems. Expanding the Societal Role of Robotics in the the Next Millennium (Cat. No. 01CH37180)*, volume 2, pages 858–864. IEEE, 2001.

[32] Roya Firoozi, Johnathan Tucker, Stephen Tian, Anirudha Majumdar, Jiankai Sun, Weiyu

Liu, Yuke Zhu, Shuran Song, Ashish Kapoor, Karol Hausman, et al. Foundation models in robotics: Applications, challenges, and the future. *arXiv preprint arXiv:2312.07843*, 2023.

[33] Xiaofeng Gao, Qiaozi Gao, Ran Gong, Kaixiang Lin, Govind Thattai, and Gaurav S Sukhatme. Dialfred: Dialogue-enabled agents for embodied instruction following. *IEEE Robotics and Automation Letters*, 7(4):10049–10056, 2022.

[34] Christophe Giovannangeli, Philippe Gaussier, and Gaël Désilles. Robust mapless outdoor vision-based navigation. In *2006 IEEE/RSJ international conference on intelligent robots and systems*, pages 3293–3300. IEEE, 2006.

[35] Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. Iqa: Visual question answering in interactive environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4089–4098, 2018.

[36] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021.

[37] Karl Moritz Hermann, Mateusz Malinowski, Piotr Mirowski, Andras Banki-Horvath, Keith Anderson, and Raia Hadsell. Learning to follow directions in street view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11773–11781, 2020.

[38] HS Hewawasam, M Yousef Ibrahim, and Gayan Kahandawa Appuhamillage. Past, present and future of path-planning algorithms for mobile robot navigation in dynamic environments. *IEEE Open Journal of the Industrial Electronics Society*, 3:353–365, 2022.

[39] Huosheng Hu and Dongbing Gu. Landmark-based navigation of industrial mobile robots. *Industrial Robot: An International Journal*, 27(6):458–467, 2000.

[40] Junning Huang, Sirui Xie, Jiankai Sun, Qiurui Ma, Chunxiao Liu, Dahua Lin, and Bolei Zhou. Learning a decision module by imitating driver's control behaviors. In *Conference on Robot Learning*, pages 1–10. PMLR, 2021.

[41] Siyuan Huang, Zhengkai Jiang, Hao Dong, Yu Qiao, Peng Gao, and Hongsheng Li. Instruct2act: Mapping multi-modality instructions to robotic actions with large language model. *arXiv preprint arXiv:2305.11176*, 2023.

[42] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International conference on machine learning*, pages 9118–9147. PMLR, 2022.

[43] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973*, 2023.

[44] Muhammad Zubair Irshad, Chih-Yao Ma, and Zsolt Kira. Hierarchical cross-modal agent for robotics vision-and-language navigation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13238–13246. IEEE, 2021.

[45] Hadi Jahanshahi and Naeimeh Najafizadeh Sari. Robot path planning algorithms: a review of theory and experiment. *arXiv preprint arXiv:1805.08137*, 2018.

[46] Sascha Jockel, Mateus Mendes, Jianwei Zhang, A Paulo Coimbra, and Manuel Crisóstomo. Robot navigation and manipulation based on a predictive associative memory. In *2009 IEEE 8th International Conference on Development and Learning*, pages 1–7. IEEE, 2009.

[47] Shirel Josef and Amir Degani. Deep reinforcement learning for safe local planning of a ground vehicle in unknown rough terrain. *IEEE Robotics and Automation Letters*, 5(4):6748–6755, 2020.

[48] Frank Joublin, Antonello Ceravola, Pavel Smirnov, Felix Ocker, Joerg Deigmoeller, Anna Belardinelli, Chao Wang, Stephan Hasler, Daniel Tanneberg, and Michael Gienger. Copal: Corrective planning of robot actions with large language models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8664–8670, 2024.

[49] Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

[50] Jon M Kleinberg. The localization problem for mobile robots. In *Proceedings 35th Annual Symposium on Foundations of Computer Science*, pages 521–531. IEEE, 1994.

[51] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, et al. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017.

[52] Yoram Koren, Johann Borenstein, et al. Potential field methods and their inherent limitations for mobile robot navigation. In *IEEE International Conference on Robotics and Automation (ICRA)*, volume 2, pages 1398–1404, 1991.

[53] Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, pages 104–120. Springer, 2020.

[54] Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. *arXiv preprint arXiv:2010.07954*, 2020.

[55] John J Leonard and Hugh F Durrant-Whyte. Mobile robot localization by tracking geometric beacons. *IEEE Transactions on robotics and Automation*, 7(3):376–382, 1991.

[56] Sergey Levine and Dhruv Shah. Learning robotic navigation from experience: principles, methods and recent results. *Philosophical Transactions of the Royal Society B*, 378(1869):20210447, 2023.

[57] Frank L Lewis and Shuzhi Sam Ge. *Autonomous mobile robots: sensing, control, decision making and applications*. CRC Press, 2018.

[58] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9493–9500. IEEE, 2023.

[59] Jinzhou Lin, Han Gao, Rongtao Xu, Changwei Wang, Li Guo, and Shibiao Xu. The development of llms for embodied navigation. *arXiv preprint arXiv:2311.00530*, 2023.

[60] Kevin Lin, Christopher Agia, Toki Migimatsu, Marco Pavone, and Jeannette Bohg. Text2motion: From natural language instructions to feasible plans. *Autonomous Robots*, 47(8):1345–1365, 2023.

[61] Li Linjun. Research and implementation of key technologies for mobile robot navigation. Master's thesis, University of Electronic Science and Technology of China, 2017.

[62] Bo Liu, Xuesu Xiao, and Peter Stone. A lifelong learning approach to mobile robot navigation. *IEEE Robotics and Automation Letters*, 6(2):1090–1096, 2021.

[63] Haokun Liu, Yaonan Zhu, Kenji Kato, Atsushi Tsukahara, Izumi Kondo, Tadayoshi Aoyama, and Yasuhisa Hasegawa. Enhancing the llm-based robot manipulation through human-robot collaboration. *arXiv preprint arXiv:2406.14097*, 2024.

[64] Jijia Liu, Chao Yu, Jiaxuan Gao, Yuqing Xie, Qingmin Liao, Yi Wu, and Yu Wang. Llm-powered hierarchical language agent for real-time human-ai coordination. *arXiv preprint arXiv:2312.15224*, 2023.

[65] Shubo Liu, Hongsheng Zhang, Yuankai Qi, Peng Wang, Yanning Zhang, and Qi Wu. Aerialvln: Vision-and-language navigation for uavs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15384–15394, 2023.

[66] Xinzhu Liu, Di Guo, Huaping Liu, and Fuchun Sun. Multi-agent embodied visual semantic navigation with scene prior knowledge. *IEEE Robotics and Automation Letters*, 7(2):3154–3161, 2022.

[67] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024.

[68] David V Lu and William D Smart. Towards more efficient navigation for robots and humans. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1707–1713. IEEE, 2013.

[69] Jinjie Mai, Jun Chen, Bing chuan Li, Guocheng Qian, Mohamed Elhoseiny, and Bernard Ghanem. Llm as a robotic brain: Unifying egocentric memory and control. *ArXiv*, abs/2304.09349, 2023.

[70] Zhao Mandi, Shreeya Jain, and Shuran Song. Roco: Dialectic multi-robot collaboration with large language models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 286–299, 2024.

[71] Gil Manor, Joseph Z Ben-Asher, and Elon Rimon. Time optimal trajectories for a mobile robot under explicit acceleration constraints. *IEEE Transactions on Aerospace and Electronic Systems*, 54(5):2220–2232, 2018.

[72] Jerry W Mao. *A Framework for LLM-based Lifelong Learning in Robot Manipulation*. PhD thesis, Massachusetts Institute of Technology, 2024.

[73] Christoforos Mavrogiannis and Ross A Knepper. Hamiltonian coordination primitives for decentralized multiagent navigation. *The International Journal of Robotics Research*, 40(10-11):1234–1254, 2021.

[74] Christoforos I Mavrogiannis and Ross A Knepper. Multi-agent path topology in support of socially competent navigation planning. *The International Journal of Robotics Research*, 38(2-3):338–356, 2019.

[75] Harsh Mehta, Yoav Artzi, Jason Baldridge, Eugene Ie, and Piotr Mirowski. Retouchdown: Adding touchdown to streetlearn as a shareable resource for language grounding tasks in street view. *arXiv preprint arXiv:2001.03671*, 2020.

[76] Mateus Mendes, A Paulo Coimbra, and Manuel M Crisostomo. Robot navigation based on view sequences stored in a sparse distributed memory. *Robotica*, 30(4):571–581, 2012.

[77] Jean-Arcady Meyer and David Filliat. Map-based navigation in mobile robots:: Ii. a review of map-learning and path-planning strategies. *Cognitive Systems Research*, 4(4):283–317, 2003.

[78] Piotr Mirowski, Andras Banki-Horvath, Keith Anderson, Denis Teplyashin, Karl Moritz Hermann, Mateusz Malinowski, Matthew Koichi Grimes, Karen Simonyan, Koray Kavukcuoglu, Andrew Zisserman, et al. The streetlearn environment and dataset. *arXiv preprint arXiv:1903.01292*, 2019.

[79] Jun Nakanishi, Shunki Itadera, Tadayoshi Aoyama, and Yasuhisa Hasegawa. Towards the development of an intuitive teleoperation system for human support robot using a vr device. *Advanced Robotics*, 34(19):1239–1253, 2020.

[80] Khanh Nguyen and Hal Daumé. Help, anna! visual navigation with natural multimodal assistance via retrospective curiosity-encouraging imitation learning. *ArXiv*, abs/1909.01871, 2019.

[81] Khanh Nguyen, Debadeepta Dey, Chris Brockett, and Bill Dolan. Vision-based navigation with language-based assistance via imitation learning with indirect intervention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12527–12537, 2019.

[82] Duy Nguyen-Tuong and Jan Peters. Model learning for robot control: a survey. *Cognitive processing*, 12:319–340, 2011.

[83] OpenAI OpenAI, Matthias Plappert, Raul Sampedro, Tao Xu, Ilge Akkaya, Vineet Kosaraju, Peter Welinder, Ruben D'Sa, Arthur Petron, Henrique P d O Pinto, et al. Asymmetric self-play for automatic goal discovery in robotic manipulation. *arXiv preprint arXiv:2101.04882*, 2021.

[84] Kevin Osanlou, Christophe Guettier, Tristan Cazenave, and Eric Jacopin. Planning and learning: A review of methods involving path-planning for autonomous vehicles. *arXiv preprint arXiv:2207.13181*, 2022.

[85] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[86] Abhishek Padalkar, Acorn Pooley, Ajinkya Jain, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anikait Singh, Anthony Brohan, et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023.

[87] Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. Teach: Task-driven embodied agents that chat. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2017–2025, 2022.

[88] Meenal Parakh, Alisha Fong, Anthony Simeonov, Tao Chen, Abhishek Gupta, and Pulkit Agrawal. Lifelong robot learning with human assisted language planners. In *CoRL 2023 Workshop on Learning Effective Abstractions for Planning (LEAP)*, 2023.

[89] Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. Reverie: Remote embodied visual referring expression in real indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9982–9991, 2020.

[90] Anis Naema Atiyah Rafai, Noraziah Adzhar, and Nor Izzati Jaini. A review on path planning and obstacle avoidance algorithms for autonomous mobile robots. *Journal of Robotics*, 2022(1):2538220, 2022.

[91] Abhinav Rajvanshi, Karan Sikka, Xiao Lin, Bhoram Lee, Han-Pang Chiu, and Alvaro Velasquez. Saynav: Grounding large language models for dynamic planning to navigation in new environments. *Proceedings of the International Conference on Automated Planning and Scheduling*, 34(1):464–474, 2024.

[92] Santhosh K Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, et al. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. *arXiv preprint arXiv:2109.08238*, 2021.

[93] Francisco Rubio, Francisco Valero, and Carlos Llopis-Albert. A review of mobile robots: Concepts, methods, theoretical framework, and applications. *International Journal of Advanced Robotic Systems*, 16(2):1729881419839596, 2019.

[94] Raphael Schumann, Wanrong Zhu, Weixi Feng, Tsu-Jui Fu, Stefan Riezler, and William Yang Wang. Velma: Verbalization embodiment of llm agents for vision and language navigation in street view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18924–18933, 2024.

[95] Dhruv Shah, Błażej Osiński, Sergey Levine, et al. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. In *Conference on robot learning*, pages 492–504. PMLR, 2023.

[96] Dhruv Shah, Ajay Sridhar, Arjun Bhorkar, Noriaki Hirose, and Sergey Levine. Gnm: A general navigation model to drive any robot. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7226–7233. IEEE, 2023.

[97] Manasi Sharma. Exploring and improving the spatial reasoning abilities of large language models. In *I Can't Believe It's Not Better Workshop: Failure Modes in the Age of Foundation Models*, 2023.

[98] Chak Lam Shek, Xiyang Wu, Dinesh Manocha, Pratap Tokekar, and Amrit Singh Bedi. Lancar: Leveraging language for context-aware robot locomotion in unstructured environments. *arXiv preprint arXiv:2310.00481*, 2023.

[99] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10740–10749, 2020.

[100] Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. Progprompt: Generating situated robot task plans using large language models. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11523–11530. IEEE, 2023.

[101] Jiankai Sun, Hao Sun, Tian Han, and Bolei Zhou. Neuro-symbolic program search for autonomous driving decision module design. In *Conference on Robot Learning*, pages 21–30. PMLR, 2021.

[102] PE Teleweck and B Chandrasekaran. Path planning algorithms and their use in robotic navigation systems. In *Journal of Physics: Conference Series*, volume 1207, page 012018. IOP Publishing, 2019.

[103] Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. Vision-and-dialog navigation. In *Conference on Robot Learning*, pages 394–406. PMLR, 2020.

[104] Sebastian Thrun and Arno Bücken. Integrating grid-based and topological maps for mobile robot navigation. In *Proceedings of the national conference on artificial intelligence*, pages 944–951, 1996.

[105] Sebastian Thrun, Wolfram Burgard, and Dieter Fox. A probabilistic approach to concurrent mapping and localization for mobile robots. *Autonomous Robots*, 5:253–271, 1998.

[106] Edmond Tong, Anthony Opipari, Stanley Lewis, Zhen Zeng, and Odest Chadwicke Jenkins. Oval-prompt: Open-vocabulary affordance localization for robot manipulation through llm affordance-grounding. *arXiv preprint arXiv:2404.11000*, 2024.

[107] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open

foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[108] Pete Trautman, Jeremy Ma, Richard M Murray, and Andreas Krause. Robot navigation in dense human crowds: Statistical models and experimental studies of human–robot cooperation. *The International Journal of Robotics Research*, 34(3):335–356, 2015.

[109] Spyros G Tzafestas. Mobile robot control and navigation: A global overview. *Journal of Intelligent & Robotic Systems*, 91:35–58, 2018.

[110] Arun Balajee Vasudevan, Dengxin Dai, and Luc Van Gool. Talk2nav: Long-range vision-and-language navigation with dual attention and spatial memory. *International Journal of Computer Vision*, 129:246–266, 2021.

[111] Sai H Vemprala, Rogerio Bonatti, Arthur Bucker, and Ashish Kapoor. Chatgpt for robotics: Design principles and model abilities. *IEEE Access*, 2024.

[112] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.

[113] Jiaqi Wang, Zihao Wu, Yiwei Li, Hanqi Jiang, Peng Shu, Enze Shi, Huawen Hu, Chong Ma, Yiheng Liu, Xuhui Wang, et al. Large language models for robotics: Opportunities, challenges, and perspectives. *arXiv preprint arXiv:2401.04334*, 2024.

[114] Jun Wang, Guocheng He, and Yiannis Kantaros. Safe task planning for language-instructed multi-robot systems using conformal prediction. *arXiv preprint arXiv:2402.15368*, 2024.

[115] Weizheng Wang, Le Mao, Ruiqi Wang, and Byung-Cheol Min. Srlm: Human-in-loop interactive social robot navigation with large language model and deep reinforcement learning. *arXiv preprint arXiv:2403.15648*, 2024.

[116] Weizheng Wang, Ruiqi Wang, Le Mao, and Byung-Cheol Min. Navistar: Socially aware robot navigation with hybrid spatio-temporal graph transformer and preference learning. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11348–11355. IEEE, 2023.

[117] Yen-Jen Wang, Bike Zhang, Jianyu Chen, and Koushil Sreenath. Prompt a robot to walk with large language models. *arXiv preprint arXiv:2309.09969*, 2023.

[118] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[119] Jiajing Wu, Jieli Liu, Yijing Zhao, and Zibin Zheng. Analysis of cryptocurrency transactions from a network perspective: An overview. *Journal of Network and Computer Applications*, 190:103139, 2021.

[120] Yi Wu, Yuxin Wu, Georgia Gkioxari, and Yuandong Tian. Building generalizable agents with a realistic and rich 3d environment. *arXiv preprint arXiv:1801.02209*, 2018.

[121] Yuchen Wu, Pengcheng Zhang, Meiying Gu, Jin Zheng, and Xiao Bai. Embodied navigation with multi-modal information: A survey from tasks to methodology. *Information Fusion*, page 102532, 2024.

[122] Fei Xia, Amir R Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: Real-world perception for embodied agents. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9068–9079, 2018.

[123] Xuesu Xiao, Bo Liu, Garrett Warnell, and Peter Stone. Motion planning and control for mobile robot navigation using machine learning: a survey. *Autonomous Robots*, 46(5):569–597, 2022.

[124] Xuesu Xiao, Zizhao Wang, Zifan Xu, Bo Liu, Garrett Warnell, Gauraang Dhamankar, Anirudh Nair, and Peter Stone. Appl: Adaptive planner parameter learning. *Robotics and Autonomous Systems*, 154:104132, 2022.

[125] Karmesh Yadav, Ram Ramrakhya, Santhosh Kumar Ramakrishnan, Theo Gervet, John Turner, Aaron Gokaslan, Noah Maestre, Angel Xuan Chang, Dhruv Batra, Manolis Savva, et al. Habitat-matterport 3d semantics dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4927–4936, 2023.

[126] An Yan, Xin Eric Wang, Jiangtao Feng, Lei Li, and William Yang Wang. Cross-lingual vision-language navigation. *arXiv preprint arXiv:1910.11301*, 2019.

[127] Shu Yang, Muhammad Asif Ali, Cheng-Long Wang, Lijie Hu, and Di Wang. Moral: Moe augmented lora for llms' lifelong learning. *ArXiv*, abs/2402.11260, 2024.

[128] Kai Ye, Siyan Dong, Qingnan Fan, He Wang, Li Yi, Fei Xia, Jue Wang, and Baoquan Chen. Multi-robot active mapping via neural bipartite graph matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14839–14848, 2022.

[129] Lance Ying, Kunal Jha, Shivam Aarya, Joshua B Tenenbaum, Antonio Torralba, and Tianmin Shu. Goma: Proactive embodied cooperative communication via goal-oriented mental alignment. *arXiv preprint arXiv:2403.11075*, 2024.

[130] Bangguo Yu, Hamidreza Kasaei, and Ming Cao. Co-navgpt: Multi-robot cooperative visual semantic navigation using large language models. *arXiv preprint arXiv:2310.07937*, 2023.

[131] Chao Yu, Xinyi Yang, Jiaxuan Gao, Jiayu Chen, Yunfei Li, Jijia Liu, Yunfei Xiang, Ruixin Huang, Huazhong Yang, Yi Wu, et al. Asynchronous multi-agent reinforcement learning for efficient real-time multi-robot cooperative exploration. *arXiv preprint arXiv:2301.03398*, 2023.

[132] Chao Yu, Xinyi Yang, Jiaxuan Gao, Huazhong Yang, Yu Wang, and Yi Wu. Learning efficient multi-agent cooperative visual exploration. In *European Conference on Computer Vision*, pages 497–515. Springer, 2022.

[133] Fengpei Yuan, Marie Boltz, Dania Bilal, Ying-Ling Jao, Monica Crane, Joshua Duzan, Abdurhman Bahour, and Xiaopeng Zhao. Cognitive exercise for persons with alzheimer's disease and related dementia using a social robot. *IEEE Transactions on Robotics*, 39(4):3332–3346, 2023.

[134] Jinsheng Yuan, Wei Guo, Fusheng Zha, Pengfei Wang, Mantian Li, and Lining Sun. A bionic spatial cognition model and method for robots based on the hippocampus mechanism. *Frontiers in Neurorobotics*, 15:769829, 2022.

[135] Fanlong Zeng, Wensheng Gan, Yongheng Wang, Ning Liu, and Philip S Yu. Large language models for robotics: A survey. *arXiv preprint arXiv:2311.07226*, 2023.

[136] K.R. Zentner, Ryan Julian, Brian Ichter, and Gaurav S. Sukhatme. Conditionally combining robot skills using large language models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 14046–14053, 2024.

[137] Bin Zhang, Hangyu Mao, Jingqing Ruan, Ying Wen, Yang Li, Shao Zhang, Zhiwei Xu, Dapeng Li, Ziyue Li, Rui Zhao, et al. Controlling large language model-based agents for large-scale decision-making: An actor-critic approach. *arXiv preprint arXiv:2311.13884*, 2023.

[138] Hongxin Zhang, Weihua Du, Jiaming Shan, Qinhong Zhou, Yilun Du, Joshua B Tenenbaum, Tianmin Shu, and Chuang Gan. Building cooperative embodied agents modularly

with large language models. *arXiv preprint arXiv:2307.02485*, 2023.

[139] Tianyao Zhang, Xiaoguang Hu, Jin Xiao, and Guofeng Zhang. A survey of visual navigation: From geometry to embodied ai. *Engineering Applications of Artificial Intelligence*, 114:105036, 2022.

[140] Xufeng Zhao, Mengdi Li, Cornelius Weber, Muhammad Burhan Hafez, and Stefan Wermter. Chat with the environment: Interactive multimodal perception using large language models. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3590–3596. IEEE, 2023.

[141] Zhonghan Zhao, Kewei Chen, Dongxu Guo, Wenhao Chai, Tian Ye, Yanting Zhang, and Gaoang Wang. Hierarchical auto-organizing system for open-ended multi-agent navigation. *arXiv preprint arXiv:2403.08282*, 2024.

[142] Sipeng Zheng, Yicheng Feng, Zongqing Lu, et al. Steve-eye: Equipping llm-based embodied agents with visual perception in open worlds. In *The Twelfth International Conference on Learning Representations*, 2023.

[143] Gengze Zhou, Yicong Hong, and Qi Wu. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7641–7649, 2024.

[144] Kaiwen Zhou, Kaizhi Zheng, Connor Pryor, Yilin Shen, Hongxia Jin, Lise Getoor, and Xin Eric Wang. Esc: Exploration with soft commonsense constraints for zero-shot object navigation. In *International Conference on Machine Learning*, pages 42829–42842. PMLR, 2023.

[145] Fengda Zhu, Xiwen Liang, Yi Zhu, Qizhi Yu, Xiaojun Chang, and Xiaodan Liang. Soon: Scenario oriented object navigation with graph-based exploration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12689–12699, 2021.

[146] Weiqin Zu, Wenbin Song, Ruiqing Chen, Ze Guo, Fanglei Sun, Zheng Tian, Wei Pan, and Jun Wang. Language and sketching: An llm-driven interactive multimodal multitask robot navigation framework. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1019–1025, 2024.